A Guide to Assessment in Veterinary Medicine

A Guide to Assessment
in
Veterinary Medical Education

Authors: Sarah Baillie, Sheena Warman and Susan Rhind

June 2014

Version 2

Table of Contents

Introduction

This document was originally developed as part of a 'Blue Sky' project entitled 'Evidence Based Development of a Common Final Examination for Veterinary Undergraduates' which was funded by the Royal College of Veterinary Surgeons Trust and published in 2008. The systematic review which underpinned the work is described in Rhind et al. (2008). Since this work was published, there has been a general increase in interest in assessment methods within the veterinary education community and many relevant developments in medical education. Hence this new version includes updates and references from studies published in the interim and an expanded 'Headlines' section providing an overview of concepts and terminology relevant to assessment.

Author Background and Contacts

Sarah Baillie BVSc, CertCHP, MSc(IT), PhD, MRCVS
Professor of Veterinary Education and Veterinary Programme Director
School of Veterinary Sciences, University of Bristol, Langford, Bristol BS40 5DU
sarah.baillie@bristol.ac.uk

Sheena Warman BSc, BVMS, DSAM, DipECVIM-CA, FHEA, MRCVS
Senior Clinical Fellow in Small Animal Medicine
School of Veterinary Sciences, University of Bristol, Langford, Bristol BS40 5DU
Sheena.Warman@bristol.ac.uk

Susan Rhind BVMS, PhD, FRCPath, PFHEA, MRCVS
Chair of Veterinary Medical Education and Director of the Veterinary Medical Education Division
Royal (Dick) School of Veterinary Studies, University of Edinburgh
susan.rhind@ed.ac.uk

Principles of Assessment

Models of the Development of Competence

One of the commonest cited models relating to assessment in medical education is that of Miller's Pyramid originally described by Miller in 1990 (Figure 1). This is a conceptual model which encompasses the elements required for clinical competence – from the underpinning cognitive levels of knowledge and application of knowledge (Knows and Knows How) to the behavioural levels of practical competence (Shows) and how a doctor (or veterinarian) actually performs in practice (Does) (Miller, 1990).

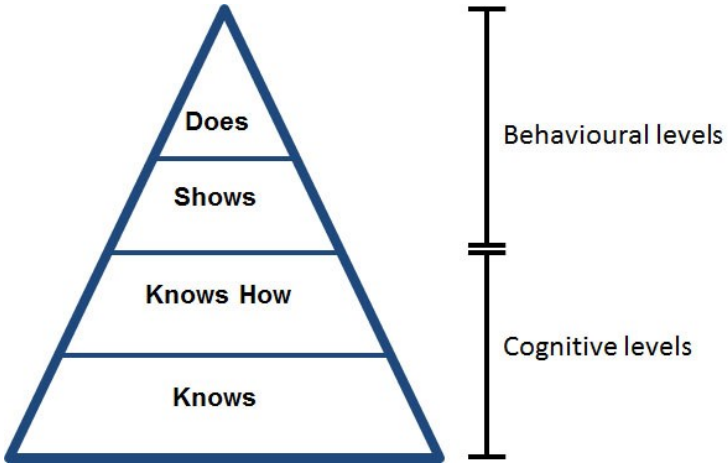


Figure 1. Miller's Pyramid

Miller's pyramid can be considered a 'condensed' version of Bloom's taxonomy (Bloom, 1984). This taxonomy divides learning into 3 domains – cognitive (knowledge based), psychomotor (skills) and affective (attitudinal). The lower 2 levels of Miller's pyramid (Knows and Knows How) map to the sequence of 6 hierarchical categories in the cognitive domain (knowledge, comprehension, application, analysis, synthesis and evaluation/judgement) reflecting a progressive contextualization of knowledge as one 'climbs' the pyramid (Figure 2). The dotted line also acknowledges that the cognitive domain of Bloom's taxonomy underpins activity at the practical levels of Miller's pyramid, especially at the 'Does' level.

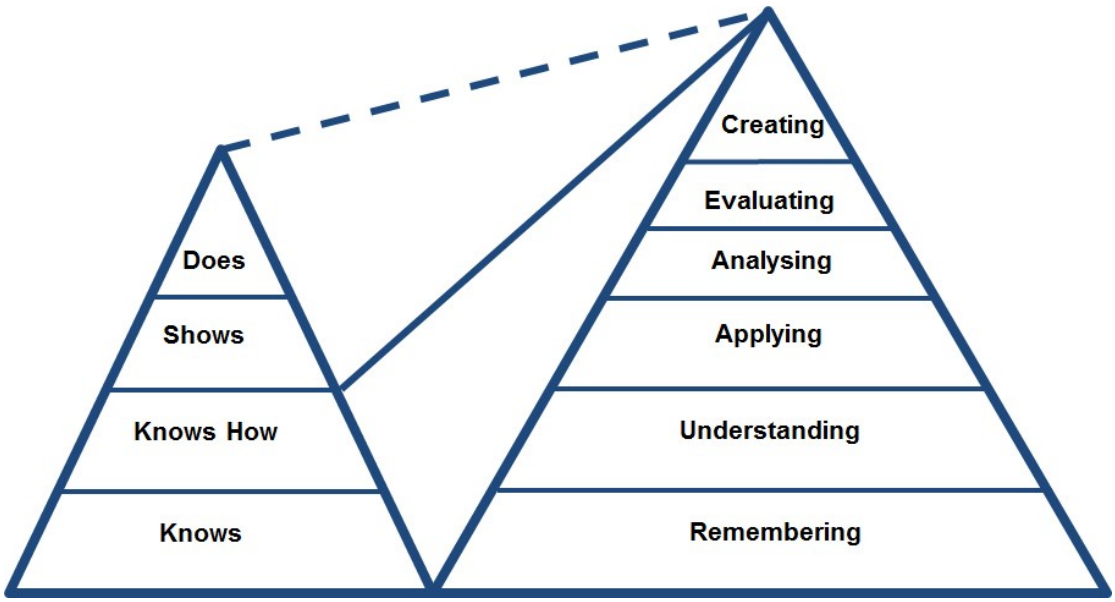


Figure 2. Miller's Pyramid and Bloom's Taxonomy Combined

Although veterinary medicine has rightly learnt much from developments in medical education, not least in assessment, it is also relevant to emphasise some key differences between our disciplines. When we consider the types of activities that veterinary graduates would be expected to undertake soon after graduating, they are much broader in scope (cross species) and in depth (e.g. surgical and imaging procedures) which has implications for our assessment at the behavioural ('Does') level of Miller's Pyramid. The most obvious manifestation of this is the current trend for assessment methods originally used in postgraduate medical education to be used in the final stages of assessment of veterinary medical students' competence.

Assessment Tools and Terminology:
The guide will present an update on key principles of assessment and assessment tools as they map to the different levels of Miller's pyramid and present current issues relating to their use. We note that throughout the international assessment literature there are potential areas of confusion resulting from discipline specific or local use of terminology to describe certain types of assessment. It is therefore a further aim of this guide to provide some clarity in assessment terminology.

Section 1 describes specific assessment methods as they map to Miller's pyramid and outlines practical issues relating to their use. Section 2 provides updated summaries of key topics linked, or directly involved, in assessment.
Throughout Section 1, we refer to the terms reliability and validity. More detail on validity is given on page 27, however in summary:

Reliability is defined as the reproducibility and accuracy of results – in assessment science, this is often computed as a reliability coefficient between 0 and 1. Two commonly reported measures of reliability are Cronbach's alpha and KR20 (Kuder–Richardson Formula 20).
Reliability is now considered an important contributor to validity in that it determines the upper limit of the validity of any test.

Validity addresses the question of whether a test measures what it is supposed to measure. Validity has been considered as one of the characteristics of assessment

instruments (van der Vleuten, 1996) with an ‘assessment formula’ presented which emphasises the other key factors that need to be considered i.e.

Utility of an assessment = reliability x validity x educational impact x acceptability x cost.

We will discuss in Section 2 how modern concepts of validity are more detailed and encompassing such that in effect validity replaces utility on the left hand side of this equation. Regardless however, this formula neatly encapsulates the many factors which can influence decisions on assessment and also serves to highlight why decisions regarding programmes of assessment are heavily influenced by local context.

Blueprinting:
One aspect of validity which we highlight in this introduction to aid the considerations presented in Section 1 is that of content validity (often referred to as blueprinting). Blueprinting refers to ensuring that the assessment is a true reflection of the taught content. It can be performed simply using a spreadsheet to map assessment questions to the course content on a pro rata basis or using more complex curriculum mapping tools.

A Programmatic View on Assessment:
Although Section 1 focuses on individual assessment methods, it is important to bear in mind that developing and describing an overall programme of assessment is an extremely useful if not increasingly essential element of assessment development and planning. Such programmes should clearly describe how assessment maps to relevant competency domains across the curriculum and capture the progressive development of the professional student.

References and Further Reading

Anderson, L. and Krathwohl, D. (2001). *A taxonomy for learning, teaching, and assessing*. 1st ed. New York: Longman.

Bloom, B. (1984). Taxonomy of Educational Objectives. In: D. McKay, ed., *The Cognitive Domain*, 1st ed. New York: Company Inc.

Miller, G. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, 65(9), pp.63--67.

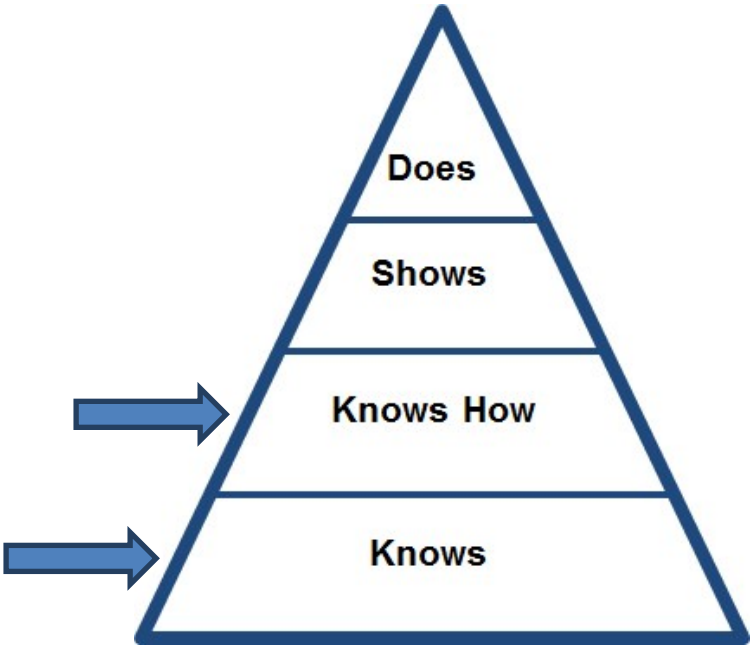
Rhind, S., Baillie, S., Brown, F., Hammick, M. and Dozier, M. (2008). Assessing Competence in Veterinary Medical Education: Where's the Evidence? *Journal of Veterinary Medical Education*, 35(3), pp.407-411.

Schuwirth, L. and Van der Vleuten, C. (2012). Programmatic assessment and Kane’s validity perspective. *Medical Education*, 46(1), pp.38--48.

Van der Vleuten, C. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*, 1(1), pp.41--67.

SECTION 1: ASSESSMENT METHODS

Miller's Pyramid 'Knows' and 'Knows How'



Focussing on the cognitive 'levels' of Miller's Pyramid, it is a reasonable aim to strive, even in fixed response questions such as MCQs, to examine at the 'Knows How' level, or using the cognitive domain of Blooms taxonomy (Figure 2), at the highest levels where possible. These structural frameworks can help focus question authors' minds on the thought processes they wish to examine when writing assessment questions.

Multiple Choice Questions (MCQs)

Knowledge Assessed: Depending on the question can range from Knows → Knows How levels of Miller's pyramid and remembering → evaluating levels of Bloom's taxonomy.

Description: MCQs are the most common written test at all levels of medical education. The most commonly used template of MCQs consists of a lead-in question or statement (stem) followed by a list of options (usually five) from which the examinee selects one answer. At the most basic level, only one of the options is correct. At higher levels, examinees are asked to choose the 'best answer', with several options being potentially correct but one being a better match to the stem than the others (by a clear margin). This type of MCQ is called the Single Best Answer type or SBA. There are other types of MCQ format including True/False, sentence completion, assertion reasoning and matching questions, but these are no longer recommended for a number of reasons including the chance of answering correctly by guessing, and the tendency for the formats to test at the lower levels of Miller's pyramid and Bloom's taxonomy. These formats have largely been replaced by SBAs, and are consigned to the 'graveyard' in Case & Swanson's guide. MCQs are used to test knowledge (factual recall) objectively and efficiently (computer-marked). MCQs can be structured to test higher order skills and levels of cognition such as understanding, application of knowledge and evaluation of information, when the question stem may take the form of a clinical vignette. The tests can be used formatively (in-training) as an indicator of progress, as well as summatively. MCQs are extensively used in veterinary assessments. One example of a computer-based, large (360 question), high stakes MCQ is the NAVLE (North American Veterinary Licensing Examination). MCQs (along with EMQs and SAQs) are used by some medical schools for 'progress testing' - a longitudinal exam with regular sampling throughout the course. The improvement in students' scores can be used to monitor progress. The MCQ exam can be presented in a paper-based format or on a computer. Computer marking results in considerable savings in staff marking time compared with marking of free text responses. For a given amount of time, MCQs give better (wider) coverage of the examinee's knowledge of a subject area than other methods e.g. essays. A robust standard-setting process, whilst time-consuming, is essential.

Considerations:
Question Format. The MCQ format may encourage students to take a superficial approach to learning if a correct answer depends purely on factual recall rather than understanding. For better authenticity in terms of testing clinical competence, it is preferable for SBAs to be based on clinical vignettes, requiring candidates to use their knowledge base to make a diagnosis, or choose an appropriate investigation or treatment, thus engaging higher-order thinking (clinical decision making). The development of the large number of good quality test items required for an exam is time consuming.
Cueing. In MCQs, and similar exam formats, cueing effects can mean that examinees are able to eliminate wrong answers and recognise the correct answer, rather than needing to work out the answer. Questions should be designed to avoid cueing. Guidance on good MCQ question writing and how to avoid some of the common pitfalls is provided in 'Case and Swanson' (see 'References' below).
'Good Practice'. When writing questions, the first thing to do is establish the "testing point"; precisely which bit of knowledge or skill are you testing? The question must be clear, and not contain superfluous information. In most cases it should be possible to arrive at an answer without looking at the options (the "cover-up test"). All distracters (i.e. incorrect or unlikely options) should be homogeneous (e.g. all are muscles,

diagnoses, drugs, etc.); plausible and attractive to the uninformed; similar to the correct answer in construction and length; and grammatically consistent and logically compatible with the stem. Try to avoid negatively phrased questions e.g. “which of the following statements is NOT TRUE” or “each of the following statements is correct EXCEPT”; this style of question inevitably fails the “cover-up test” and should only be used when there is no other way of addressing the testing point of the question.

The Test-Wise Student. There are a number of ways a test-wise student can gain an advantage based on the way MCQs are written. Some tips include: Avoid grammatical cues e.g. do all the answer options follow grammatically from the question? Avoid absolute terms such as “always” or “never” in answer options *(these are unlikely to be the correct answer and are ruled out by the test-wise student)*. Avoid vague terms in the answer options e.g. “rarely”, “usually”. Is the correct answer obviously different to the rest i.e. correct answer is longer, more specific, or more complete than other options? All answer options should be of similar length. Avoid word repeats, where a word or phrase is included in the question (stem) and in the correct answer. Beware convergence strategy where the correct answer includes the most elements in common with the other options. Avoid logical cues when a subset of the options is ‘collectively exhaustive’.

Negative Marking. Scoring of MCQs is traditionally 1 = correct answer and 0 = incorrect answer, but there is a possibility to use negative marking, where the correct answer gains a mark, the wrong answer loses a mark and no response scores zero. Negatively marked MCQs are known to be stressful and affected by the student’s willingness to ‘take a risk’.

Reliability: The reliability should be monitored with a target coefficient (Cronbach's alpha) in excess of 0.7 - 0.8

Key Points:

- High reliability
- Computer marking saves time and resources
- Writing items to test higher cognitive levels is time consuming
- Feedback often limited to overall score or score in different sections (due to question security)
- Easy to blueprint comprehensively
- Requires significant staff training and quality assurance
- Standard setting is time consuming

References and Further Reading

A North American Study of the Entry-Level Veterinary Practitioner. (2010). 1st ed. [ebook] Bismarck: National Board of Veterinary Medical Examiners. Available at: http://www.nbvme.org/image/cache/2010_NAVLE_job_analysis_report.pdf [Accessed 6 May 2014].

Anderson, J. (2004). Multiple choice questions revisited. *Medical Teacher*, 26(2), pp.110--113.

Case, S. and Swanson, D. (2002). Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd ed. [ebook] Philadelphia: National Board of Medical Examiners. Available at: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf [Accessed 6 May. 2014].

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher*, 26(8), pp.709--712.

Tractenberg, R., Gushta, M., Mulroney, S. and Weissinger, P. (2013). Multiple choice questions can be designed or revised to challenge learners’ critical thinking. *Advances in Health Sciences Education*, 18(5), pp.945—961

Extended Matching Questions (EMQs)

Knowledge Assessed: Depending on the question can range from Knows → Knows How levels of Miller's pyramid and remembering → evaluating levels of Bloom's taxonomy.

Description: The EMQ format has four components and starts with a title or theme statement defining the subject area e.g. 'Equine Surgery - Colic'. The title is followed by the list of 'options' (numbered or lettered) - the possible answers to the question/s or 'item/s' that follow. A lead in statement then provides instructions and links the list of answers (options) to the question/s (item/s), which often take the form of clinical vignettes. The examinee has to respond to each question by selecting the best answer from a large list (range from 5 up to 20+), where one or more answers are potentially correct. Where there are several questions under one title, each answer can be used once, more than once or not at all. Ordering the list of answers alphabetically helps to minimise cueing. Usually 1 to 2 minutes is allowed per question.

Considerations: as for MCQs

Reliability: The reliability should be monitored with a target coefficient (Cronbach's alpha) in excess of 0.7 - 0.8

Key Points:

- Potentially high reliability
- Writing items to test higher cognitive levels is time consuming
- Linked items can reduce the choice of topics and therefore reduce sampling
- Feedback often limited to overall score or score in different sections (due to question security)

References and Further Reading

Beullens, J., Damme, B., Jaspaert, H. and Janssen, P. (2002). Are extended-matching multiple-choice items appropriate for a final test in medical education? *Medical Teacher*, 24(4), pp.390--395.

Case, S. and Swanson, D. (2002). *Constructing Written Test Questions for the Basic and Clinical Sciences*. 3rd ed. [ebook] Philadelphia: National Board of Medical Examiners. Available at: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf [Accessed 6 May. 2014].

van Bruggen, L., Manrique-van Woudenbergh, M., Spierenburg, E. and Vos, J. (2012). Preferred question types for computer-based assessment of clinical reasoning: a literature study. *Perspectives on Medical Education*, 1(4), pp.162--171.

Short-Answer Questions (SAQs)

Knowledge Assessed: Depending on the question can range from Knows → Knows How levels of Miller's pyramid and remembering → creating levels of Bloom's taxonomy.

Description: A written test consisting of a series of questions that require students to supply or formulate an answer rather than choose from a list of options (as in MCQs). The answer format is quite heterogeneous. At one end of the spectrum a short and quite specific answer is required e.g. one word (fill in the blank) or completion of a sentence. Alternatively, a SAQ may require the examinee to construct a short response (several sentences, a plan or a diagram) and in some contexts write a short version of an essay. Questioning can be directed to test a specific objective or area. The question format may be based on a case scenario or set of data and may include additional information e.g. images. Sometimes several SAQs are written as a linked series covering a particular topic area. Compared to MCQ/EMQ, there is no cueing effect as examinees are not presented with the correct answer amongst a number of other choices.

Considerations: Considerable resources required for marking - mainly done 'by hand', although computer marking can be used for single word and short phrase answers. Basic factual knowledge is generally more efficiently examined using computer-based / computer-marked alternatives (MCQs/EMQs). Compared with essays, SAQs are easier to write and mark and have the potential to be more objective although questions need to be worded carefully to elicit the desired answer. In linked SAQs, question design should ensure the examinee's progression through the answer is not blocked by an incorrect response early on.

Reliability: Reliability is affected by marker subjectivity with regard to what constitutes an acceptable answer, which is more of a problem the longer and less structured the answer format. To improve reliability, outline answers, marking schemes and double-marking should be employed.

Key Points:

- Resource intensive marking compared to MCQ/EMQ (unless computer-markable)
- Heterogeneity in interpretation of the term
- Reliability improved if structured marking schemes, clear outline answers and independent double scoring employed
- No cueing effect
- Provision of written feedback possible but time consuming

References and Further Reading

Rademakers, J., ten Cate, T. and Bar, P. (2005). Progress testing with short answer questions. *Medical Teacher*, 27(7), pp.578--582.

Schuwirth, L. and Van der Vleuten, C. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, 38(9), pp.974--979.

Schuwirth, L. (2003). ABC of learning and teaching in medicine: Written assessment. *British Medical Journal*, 326(7390), pp.643-645.

Essays

Knowledge Assessed: Depending on the question can range from Knows → Knows How levels of Miller's pyramid and remembering → creating levels of Bloom's taxonomy.

Description: 'a short literary composition on a particular theme or subject, usually in prose and generally analytic, speculative, or interpretative.'[a] Sometimes also referred to as 'long answer' or 'extended answer' questions. A variation is the modified essay question, which may include e.g. an element of data handling. It should be clear to students whether the essay is being assessed / marked as a structured argument or is being used as a means of testing knowledge. For the latter, more efficient alternatives are preferable.

Considerations: Marking is labour intensive. Techniques to detect plagiarism should be considered. Not recommended for high stakes assessment.

Reliability: Reliability is low as sampling across content tends to be low unless a large number of essays are used.

Key Points:

- Resource intensive marking
- Low reliability
- Double marking recommended to improve reliability
- Heterogeneity in interpretation of the word 'essay' which can be confusing for students and make comparison as a 'method' confusing.
- Provision of written feedback possible but time consuming
- Not recommended for high stakes assessment

References and Further Reading

[a]Dictionary.com, (2014). *Dictionary.com*. [online] Available at: http://dictionary.reference.com [Accessed 25 Jan. 2014].

Schuwirth, L. and van der Vleuten, C. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, 38(9), pp.974--979.

Schuwirth, L. (2003). ABC of learning and teaching in medicine: Written assessment. *British Medical Journal*, 326(7390), pp.643-645.

Viva / Viva Voce / Oral

Knowledge Assessed: Depending on the question can range from Knows → Knows How levels of Miller's pyramid and remembering → creating levels of Bloom's taxonomy. Oral defence (a viva) is often as part of the assessment of a project or thesis. In this context, there is an important element of authenticating the work as belonging to a given student. If designed appropriately, this method can also be used to assess oral communication skills and professionalism.
Note: Vivas have been mostly replaced by more reliable and efficient assessment methods (see 'Considerations' and 'Reliability' issues below).

Description: The viva format involves the examinee being questioned by one or more examiners using an interview or discussion-like format typically to ascertain knowledge of a subject area or the ability to solve a clinical problem. This is followed by discussion and questioning aimed at probing the examinee's depth and breadth of knowledge, understanding, reasoning, and decision making process. A viva can be used to explore ethical issues, assess professionalism, attitudes and communication skills. As with several other forms of assessment, there is considerable variation in the format and use of this type of assessment.

Considerations: If used as part of routine examinations for all students, the time and resources required are considerable. This is even more of a problem when the number of questions or cases presented is increased (as one way of trying to improve reliability). Vivas (as well as other one to one encounters) can be subject to "Halo effects" i.e. the effect whereby a judgement on one aspect is influenced by an overall impression of the person or where the judgement is influenced by the performance of previous candidates in contrast to the current candidate.
Even with standardised content and structure, one student's assessment can vary markedly from another. These issues mean that the use of oral examinations in any form of high stakes assessment setting is not recommended.

Reliability: Reliability is often low due to a lack of standardisation of questioning and marking, and the possibility of examiner bias (use of favoured and / or irrelevant questions), and 'halo effects'. Reliability can be improved when using the same questions for all students (but this will require corralling to prevent later candidates being advantaged), having a structured marking system, increasing the number of vivas per examinee, having a testing time of 4 hours or more and with improved examiner training.

Key Points:

- Heterogeneity in interpretation of the method
- Low reliability unless multiple examiners, multiple cases and large testing time
- Often seen as having high authenticity to examiners
- Resource intensive
- Immediate face to face feedback can be built in

References and Further Reading

Davis, M. and Karunathilake, I. (2005). The place of the oral examination in today's assessment systems. *Medical Teacher*, 27(4), pp.294--297.

Muzzin, L. and Hart, L. (1985). Oral Examinations. In: V. Neufeld and G. Norman, ed., *Assessing Clinical Competence*, 1st ed. New York: Springer Publishing Company, pp.71-93.

Wass, V., Wakeford, R., Neighbour, R. and van der Vleuten, C. (2003). Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. *Medical Education*, 37(2), pp.126--131.

The 'Spot' Test

Knowledge Assessed: Depending on the question can range from Knows → Knows How levels of Miller's pyramid and remembering → creating levels of Bloom's taxonomy.

Description: This category is included as it has been a traditional assessment format in many UK veterinary schools – particularly in disciplines such as anatomy and pathology. However, the format is increasingly being replaced, in part or completely, by computerised assessments using high quality images. Various local terms are used to describe this type of assessment including 'Spot', 'Steeplechase', 'Timed stations' or 'Bellringer'. However, there are few references in the literature to the method and it should not be confused with methods assessing at the 'Shows' level of Miller's pyramid. The format usually has examinees moving around a series of stations consisting of e.g. a specimen, a labelled dissection or radiograph. The answer may be one word or involve a response that requires some level of deduction or diagnostic skill i.e. similar to that described under the category of short answer questions. As for SAQs, therefore, the same reliability issues exist, which can be improved using structured marking schemes.

Considerations: Resources required to set up the stations, run the exam and marking. Often the same knowledge could be tested more efficiently and reliably by using images within an MCQ or SAQ test.

Reliability: Reliability will be compromised if the number of items is small and when marking is not structured.

Key Points:

- Has been in common use but is being replaced by computerised assessment and marking (where possible)
- Little published in literature on description
- Heterogeneity in interpretation of the term
- Reliability improved if structured marking schemes employed
- Provision of written feedback possible but time consuming
- Consider using images within more reliable and evidence-based forms of assessment

References and Further Reading

Note: Literature searching to date for further information on this method has found no specific papers

Script Concordance Test (SCT)

Knowledge Assessed: Depending on the question can range from Knows → Knows How levels of Miller's pyramid and remembering → evaluating levels of Bloom's taxonomy.

Description: The Script Concordance Test (SCT) is a tool designed to assess decision-making and clinical reasoning skills. In everyday work, experienced clinicians refer to 'scripts' when using their knowledge to make decisions. These scripts are built up over years in clinical practice. The SCT investigates the organisational structure of an examinee's knowledge when presented with a situation where a decision needs to be made using information or data about a clinical case. The SCT is a written exam that starts with a clinical scenario or vignette that summarises the case. This is followed by a proposed diagnosis or suggested treatment or action. Examinees have to rate the effect of further information or findings on the probability of the diagnosis / treatment being: more certain / likely, unchanged or less certain / likely, using a 5-point scale. The answers are compared to those of a panel of experts. The marking system usually takes into account the variation in expert opinion, with answers being weighted accordingly i.e. an answer the same as the majority of experts scores highest but answers that correspond to those chosen by some experts still receive some credit. Alternatively, there is an agreed single best answer.

Considerations: The main considerations are the time and practice required in developing suitable test items. The numbers of experts required make this a challenging format to develop in veterinary medicine.

Reliability: Reliability has been found to be high if there are sufficient questions. Formulation of up to 5 questions per case has been shown to be an efficient way to optimize the reliability of SCT score.

Key Points:

- Written test based on scenario / vignette
- Significant training required for item writing
- Considerable time needed to produce each question
- Good reliability if sufficient questions are used
- Requirement for suitably qualified panels of examiners to produce scoring system
- Feedback often limited to overall score or score in different sections (due to question security)

References and Further Reading

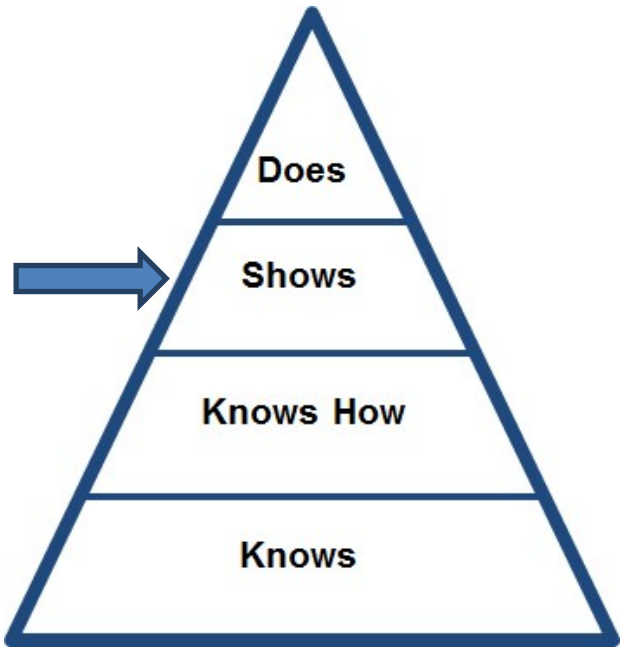
A veterinary example of a Script Concordance Test available at: http://www.veteducation.org/resources/view_workshop2006_SCT_examples.pdf [Accessed 6 May. 2014].

Charlin, B., Roy, L., Brailovsky, C., Goulet, F. and van der Vleuten, C. (2000). The Script Concordance test: a tool to assess the reflective clinician. *Teaching and Learning in Medicine*, 12(4), pp.189--195.

Gagnon, R., Charlin, B., Lambert, C., Carriere, B. and van der Vleuten, C. (2009). Script concordance testing: more cases or more questions? *Advances in Health Sciences Education*, 14(3), pp.367--375.

Meterissian, S., Zabolotny, B., Gagnon, R. and Charlin, B. (2007). Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *The American Journal of Surgery*, 193(2), pp.248--251.

Miller's Pyramid 'Shows'



The 'Shows' level of Miller's pyramid can be considered as assessing practical ability or competence at a task 'in vitro'.

Objective Structured Clinical Examination (OSCE)

Description: The Objective Structured Clinical Examination (OSCE) was introduced in medical education nearly 40 years ago as a more standardised, objective and reliable way of assessing certain clinical skills (Harden et al. 1975) and is now in widespread use. The exam consists of multiple mini-stations (typically 10 – 20) which the examinees rotate round in sequence, completing a variety of tasks. Each station in the circuit lasts the same amount of time; from about 5 – 6 minutes for basic practical skills e.g. gloving, up to 20 minutes when embracing multiple aspects of a patient interaction e.g. history taking, physical examination, diagnosis and treatment plan. The examinee reads the scenario, then enters the station and undertakes the task. The OSCE is now widely adopted in veterinary education and typical stations test e.g. placing a muzzle on a dog, bandaging a wound, placement of an intravenous catheter, preparing a blood smear, history taking (gathering information from a client). The station set-up varies and can include: live animals, models, part-task trainers, laboratory equipment, and simulated clients. The selection of stations should be representative of, and mapped (blueprinted) to, the taught course. With the more holistic OSCE (15 – 20 minutes patient interactions) blueprinting needs to consider several dimensions of competence within each station including: stages in a clinical case, body systems and, in veterinary medicine, species. An example of an OSCE can be found on the RCVS website (RCVS Awards, 2014).

Marking: Detailed Checklists
Originally, OSCEs were marked using a very detailed checklist often with 15 – 25 items that the examinee did or did not complete / undertake. Each item can be equally weighted i.e. 1 or 0 although some critical steps (e.g. fatal errors, a break in sterility, etc.) may carry a heavier weighting (more marks) or be a requirement to pass the station. The checklists are usually accompanied by a global rating scale for the examiner to make a more subjective judgement (selecting one of 4 - 7 categories with descriptors across the spectrum from a bad fail to an excellent pass). The pass mark is usually calculated via a borderline regression or borderline group method using both the global rating and the checklist score (Boursicot et al. 2007).

Marking: Global Rating Scales (GRS)
Traditionally, detailed checklists were considered to be more objective and reliable than global rating scales; however research has challenged this view and there is evidence that GRS are more reliable and able to measure increasing levels of expertise (Cunnington et al. 1996, Regehr et al. 1998) Thus, in recent years, GRS have grown in popularity.

Skills Assessed: Clinical practical, technical & diagnostic skills, treatment planning, and communication skills.

Considerations: Considerable resources are required (costs of equipment, consumables, and personnel / staff time) to develop and set up the stations, and to run the OSCE. However, the checklists and rating scales can be computer marked. When developing OSCE stations a team is required to write the scenarios and itemised checklists. Examiner training and briefing sessions are also important.

Reliability: Reliability is usually high if there are enough sampling (around 15 - 20 stations). However, examiners need to be trained and station and inter-rater (examiner) reliability should be monitored. The exam is fair and objective as the same scenarios are presented to all examinees and the same marking criteria are applied.

Key Points:

- High reliability compared to a few long cases of individual clinical examinations
- Potential to compromise validity by excessively deconstructing tasks
- Resource intensive to establish, set up and run
- Can provide detailed specific feedback

References and Further Reading

Boursicot, K., Roberts, T. and Pell, G. (2007). Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education*, 41(11), pp.1024--1031.

Cunnington, J., Neville, A. and Norman, G. (1996). The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education*, 1(3), pp.227--233.

Davis, M., Ponnamperuma, G., McAleer, S. and Dale, V. (2006). The objective structured clinical examination (OSCE) as a determinant of veterinary clinical skills. *Journal of Veterinary Medical Education*, 33(4), pp.578--587.

Harden, R., Stevenson, M., Downie, W. and Wilson, G. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1(5955), p.447.

Hecker, K., Read, E., Vallev, Krebs, G., Donszelmann, D., Muelling, C. and Freeman, S. (2010). Assessment of First-Year Veterinary Students' Clinical Skills Using Objective Structured Clinical Examinations. *Journal of Veterinary Medical Education*, 37(4), pp.395--402.

Hodges, B. and McIlroy, J. (2003). Analytic global OSCE ratings are sensitive to level of training. *Medical Education*, 37(11), pp.1012--1016.

Hodges, B. (2006). The objective structured clinical examination: Three decades of development. *Journal of Veterinary Medical Education*, 33(4), pp.571--577.

Ma, I., Zalunardo, N., Pachev, G., Beran, T., Brown, M., Hatala, R. and McLaughlin, K. (2012). Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. *Advances in Health Sciences Education*, 17(4), pp.457--470.

May, S. and Head, S. (2010). Assessment of technical skills: best practices. *Journal of Veterinary Medical Education*, 37(3), pp.258--265.

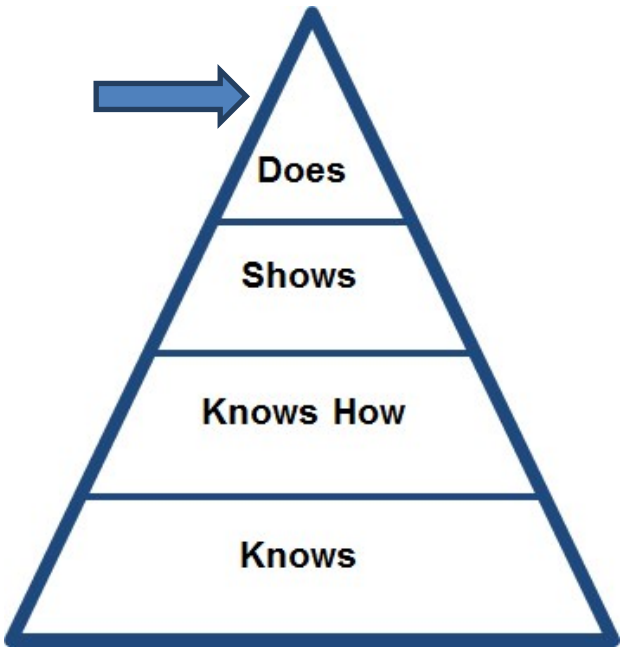
Norman, G., van der Vleuten, C. and De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education*, 25(2), pp.119--126.

RCVS Awards, (2014). *OSCE stations*. [online] Available at: http://awardingbody.rcvs.org.uk/examinations/practical-examinations/osce-stations/ [Accessed 8 May. 2014].

Regehr, G., MacRae, H., Reznick, R. and Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, 73(9), pp.993--997.

Van der Vleuten, C., Norman, G. and De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education*, 25(2), pp.110--118.

Miller's Pyramid 'Does'



Assessment at the top level of Miller's pyramid is often seen as the holy grail of clinical assessment. In contrast to performance assessment 'in vitro' discussed in the previous section, assessment at this level can be considered as performance assessment 'in vivo' i.e. in the workplace.
It is important to recognise that there is no one preferred method of assessing the professionalism which is an essential element of assessing performance at this level of the pyramid. Indeed, 9 different clusters of assessment tools have been described in medical education that have relevance for assessment in this area (Wilkinson et al. 2009). It follows from this that the use of multiple methods is desirable to allow 'triangulation' of information looking at different aspect of professional behaviour (van Mook et al. 2009).

References and Further Reading

van Mook, W., van Luijk, S., O'Sullivan, H., Wass, V., Schuwirth, L. and van der Vleuten, C. (2009). General considerations regarding assessment of professional behaviour. *European Journal of Internal Medicine*, 20(4), pp.90--95.

Wilkinson, T., Wade, W. and Knock, L. (2009). A blueprint to assess professionalism: results of a systematic review. *Academic Medicine*, 84(5), pp.551--558.

Mini-clinical Evaluation Exercise (mini-CEX)

Description: The mini-CEX involves direct observation of a trainee by one examiner during a clinical encounter with a real patient in the normal work setting e.g. on a ward or in an out-patient clinic. The mini-CEX evolved from the original clinical evaluation exercise (CEX) which was developed to replace orals used in the assessment of clinical competency. The CEX is no longer used since its focus on a relatively long (typically 2 hour) pre-planned single patient encounter in a clinical setting immediately causes problems in terms of assessment reliability.
In the mini-CEX, the observation lasts 15 – 20 minutes and is followed by immediate feedback from the examiner. Typically, multiple mini-CEXs are used in a variety of situations. The observation is marked using a standardised tick box form that is used to record information about the case, setting, trainee and examiner (for an example of a marking sheet see: Norcini, 2005). Performance is rated for a list of skills as: at, above or below expectation. Primarily used formatively with feedback and an action plan is produced which is structured to support the trainee's learning.

Mini-CEX forms may be included in a trainee's portfolio. A veterinary version, the Veterinary Clinical Assessment Tool (V-CAT), based on the mini-CEX, has been developed and trialled at the Faculty of Veterinary Medicine, University of Glasgow.

Skills Assessed: History taking, physical examination, problem solving, clinical reasoning, communication.

Considerations: With a certain amount of planning, the mini-CEX is both feasible and can be fitted into routine clinical training and is of educational benefit. The patient/s chosen should be typical of the trainee's case load.

Reliability: Reliability increases with the number of encounters i.e. mini-CEXs with 6– 8 giving acceptable reliability. Assessor training is also important for reliability and to improve the quality of feedback.

Key Points:

- High authenticity
- Reliability increases with number of examinations (mini-CEXs) performed

References and Further Reading

Kogan, J., Bellini, L. and Shea, J. (2002). Implementation of the Mini-CEX to Evaluate Medical Students' Clinical Skills. *Academic Medicine*, 77(11), pp.1156--1157.

Norcini, J. (2005). The Mini Clinical Evaluation Exercise (mini-CEX). *The Clinical Teacher*, 2(1), pp.25--30.

Setna, Z., Jha, V., Boursicot, K. and Roberts, T. (2010). Evaluating the utility of workplace-based assessment tools for speciality training. *Best Practice & Research. Clinical Obstetrics & Gynaecology*, 24(6), pp.767--782.

Directly Observed Procedural Skills (DOPS)

Description: Directly Observed Procedural Skills (DOPS), also referred to as Direct Observation of Procedural Skills, are designed specifically to assess practical skills in a workplace setting. A trainee is observed and scored by an assessor while performing a routine practical procedure during his / her normal clinical work. The assessor uses a standard DOPS form to score the technique (for an example of a DOPS form see: Wilkinson et al. 2008). For any particular skill the trainee has to pass a number of repeated assessments (typically six) to be signed off as competent at that skill. However, in veterinary medicine it is often the students' responsibility to request assessment of a skill when they judge that they have developed their competency to the required level; in this context retrospective assessment is not appropriate.

Skills Assessed: Practical / technical.

Considerations: DOPS are run during normal clinical work and, with a certain amount of planning and organisation, this represents a feasible way of assessing the key procedures and practical skills required for particular disciplines / specialties.

Reliability: Use in medical specialties indicates that six observations i.e. DOPS exams are required per procedure for a reasonable level of reliability.

Key Points:

- High authenticity
- Multiple assessments of the same skill
- Present a valuable opportunity for formative feedback

References and Further Reading

Magnier, K., Dale, V. and Pead, M. (2012). Workplace-Based Assessment Instruments in the Health Sciences. *Journal of Veterinary Medical Education*, 39(4), pp.389--395.

McLeod, R., Mires, G. and Ker, J. (2012). Direct observed procedural skills assessment in the undergraduate setting. *The Clinical Teacher*, 9(4), pp.228--232.

Wilkinson, J., Crossley, J., Wragg, A., Mills, P., Cowan, G. and Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education*, 42(4), pp.364-373.

Wragg, A., Wade, W., Fuller, G., Cowan, G. and Mills, P. (2003). Assessing the performance of specialist registrars. *Clinical Medicine*, 3(2), pp.131—134.

360º (Multi-source Feedback)

Description: Involves collecting information about a clinician's performance in the workplace from those working with that individual. Feedback is gathered using a structured form or questionnaire (for an example of a 360º assessment form see: Wood et al. 2006). Different members of the clinical team assess the individual's performance and particularly his or her professional behaviour. Those 'assessing' the individual include staff who are more senior, more junior and peers; representatives of all groups in the clinician's daily working environment (not just co-professionals); e.g. clients The feedback is used as part of appraisals and to help clinicians gain insight into their professional development.

Skills Assessed: Communication, team working, professionalism.

Considerations: It is feasible for those working with the trainee to participate in this form of assessment as it is based on observations made during every day work. Each rater fills out a short form that takes 5 – 10 minutes to complete.

Reliability: Reliability depends on feedback from a wide enough range of team members (from all levels) and sufficient raters (usually 8 to 12). An important part of 360º is making good use of the feedback.

Key Points:

- Allows feedback from range of individuals (a variety of staff +/- patients)
- Resource intensive
- Very useful information gained about professional behaviour

References and Further Reading

Evans, R., Elwyn, G. and Edwards, A. (2004). Review of instruments for peer assessment of physicians. *British Medical Journal*, 328(7450), p.1240.

Wood, L., Hassell, A., Whitehouse, A., Bullock, A. and Wall, D. (2006). A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. *Medical Teacher*, 28(7), pp.185--191.

Wood, L., Wall, D., Bullock, A., Hassell, A., Whitehouse, A. and Campbell, I. (2006). 'Team observation': a six-year study 1 of the development and use of multi-source feedback (360-degree assessment) in obstetrics and gynaecology training in the UK. *Medical Teacher*, 28(7), pp.177--184.

Case-based Discussion

Description: A formal discussion between a trainee and an assessor about a case that the trainee has managed and been directly responsible for. Case-based discussions are primarily used for formative assessment ('in-training'). During the discussion, the trainee refers to the case records. The assessor will probe the trainee's depth of understanding, decision-making and clinical judgement. The trainee has the opportunity to talk about any issues that arose and explain decisions. The assessor can also determine the quality of various aspects of case management e.g. synthesising information, prioritising, planning and record keeping.

A structured assessment form is used to record basic case details and rate the key skill areas (for an example of an assessment form see: Intercollegiate Surgical Curriculum Website https://www.iscp.ac.uk/static/public/cbd_form.pdf). The discussion is followed by a short feedback session.

Choosing a challenging case enables the trainee to maximise the benefits of discussing and reflecting on a case with a more senior clinician. The format is broadly similar to that described for 'Chart Stimulated Recall' where a doctor's own cases are used as the basis for a structured oral examination.

Skills Assessed: Application of knowledge, decision making, clinical judgement, professionalism.

Considerations: The discussion lasts about 20 minutes with 5 to 10 minutes for feedback. Typically the assessment is performed several times per placement and over that time should cover a range of cases that are typical for the particular speciality. Although undertaken during workplace training the assessment is not carried out during a clinical encounter but in an office or meeting room setting.

Reliability: Reliability depends in part on the assessor's training in use of the form and giving feedback. However, as only one rater is involved there is potential for bias and variable reliability. Essentially as this is a structured oral it suffers from the same problems of reliability as other orals described earlier.

Key Points:

- High authenticity
- Mostly used formatively
- Low reliability

References and Further Reading

Cunnington, J., Hanna, E., Turnhbull, J., Kaigas, T. and Norman, G. (1997). Defensible assessment of the competency of the practicing physician. *Academic Medicine*, 72(1), pp.9--12.

Guidance notes for CBD. (2010). 1st ed. [ebook] The Royal College of Surgeons of England. Available at: https://www.iscp.ac.uk/static/public/cbd_guidance.pdf [Accessed 6 May. 2014].

Jennett, P. and Affleck, L. (1998). Chart audit and chart stimulated recall as methods of needs assessment in continuing professional health education. *Journal of Continuing Education in the Health Professions*, 18(3), pp.163--171.

Setna, Z., Jha, V., Boursicot, K. and Roberts, T. (2010). Evaluating the utility of workplace-based assessment tools for speciality training. *Best Practice & Research. Clinical Obstetrics & Gynaecology*, 24(6), pp.767--782.

Observation on Rotations

Description: Students are observed and assessed during clinical work i.e. on intramural and extramural rotations / clerkships. This type of assessment has sometimes been referred to as 'Longitudinal Evaluation of Performance (LEP)'. The assessment is based on performance over a period of time (days to weeks) and a number of skills can be rated from basic factual knowledge to technical skills as well as other aspects of professional behaviour. The method of marking and assigning grades varies considerably. Students are often assigned a grade at the end of the rotation / placement, which can be derived from a global rating form that includes general categories of professional and clinical ability e.g. knowledge, clinical skills, communication skills, case responsibility, preparation and professionalism.

The assessment may be undertaken by one member of staff or several members of the team. If individuals other than the clinicians are involved the assessment approaches the 360º evaluations used in medicine described earlier.

Skills Assessed: Knowledge, application of knowledge, clinical/practical skills, diagnostic skills, clinical reasoning, communication skills, attitudes and professionalism.

Considerations: As the assessment is embedded in day-to-day work there are relatively low demands on resources.

Reliability: Reliability tends to be low as the assessment often lacks standardisation e.g. observational frequency varies, marking can be very subjective as it is often based on 'clinical impressions', can be affected by 'halo effects', and inter-rater reliability is poor. Additionally, staff are sometimes reluctant to fail students. The objectivity and reliability can be improved if checklists are used and the frequency and breadth of assessment is increased.

Key Points:

- Based on observation of students
- Low reliability
- Subjective and prone to 'halo effects'
- Can provide useful opportunity for feedback

References and Further Reading

Miller, G. (1990). The assessment of clinical skills / competence / performance. *Academic Medicine*, 65(9), pp.63--7.

Prescott-Clements, L., van der Vleuten, C., Schuwirth, L., Hurst, Y. and Rennie, J. (2008). Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). *Medical Education*, 42(5), pp.488--495.

Turnbull, J. and Van Barneveld, C. (2002). Assessment of clinical performance: in-training evaluation. *Springer International Handbooks of Education*, 7, pp.793--810.

Portfolios

Description: This is a collection of work developed as a cumulative 'body of evidence' to demonstrate the student's learning and achievements. **It is not an examination format in its own right,** rather a receptacle containing a mixture of materials, each piece assessable to predefined marking criteria which may be graded or pass/fail.

Hence although included in this section on the 'Does' level of Miller's pyramid, in real terms the portfolio itself contains evidence relating to 'Does'. The content, which can be paper-based or in an electronic format (e-portfolio), is collected during day-to-day activities and is typically quite diverse e.g. written assignments, reports, feedback, case studies and projects. Supplementary material such as photographs, videos and curriculum vitae may be included.

A portfolio can also be used to plan learning needs and to monitor progress e.g. with checklists of skills and activity logs. Evidence of the student's reflections on learning is a valuable aspect of a portfolio. Portfolios have been used in veterinary nurse training in the UK for many years.

The approach to the assessment of portfolios and the criteria applied are quite variable and depend on content. Assessment is often an ongoing process, can be formative and/or summative, and in an ideal situation involves more than one marker. Interviews provide an opportunity to determine how well the portfolio reflects the student's achievements. Portfolios are not always formally assessed, instead the requirement being for the provision of evidence that certain tasks have been completed.

Skills Assessed: Knowledge, knowledge application and interpretation, case recording and interpretation, attitudes and professionalism (skills not always easy to assess using other methods).

Considerations: Staff time is a major consideration as portfolios are labour intensive to supervise and mark, and to provide feedback, although the workload may be spread throughout the year. Student perceptions of value vary from being seen as providing a useful framework for learning, to having a low return relative to the time and effort expended. Uptake and engagement vary and are affected by: learner type and maturity; tutor enthusiasm and support. Using a framework to align portfolio content with curriculum or course outcomes will help students produce a representative and comprehensive 'body of evidence'.

Reliability: Achieving reliability can be difficult and is affected by the diverse content of a portfolio and the subjective aspects of the evaluation particularly if only one examiner is involved. Reliability can be improved using rating scales and having more than one marker. Assessing the student's process of reflection is not straightforward if that is deemed desirable for a given context.

Key Points:

- Heterogeneity in meaning – covers many different formats
- Resource intensive
- Assessing reflection is difficult and controversial

References and Further Reading

Challis, M. (1999). AMEE Medical Education Guide No. 11 (revised): Portfolio-based learning and assessment in medical education. *Medical Teacher*, 21(4), pp.370--386.

Davis, M. and Ponnamperuma, G. (2005). Portfolio assessment. *Journal of Veterinary Medical Education*, 32(3), pp.279-284.

Friedman Ben-David, M., Davis, M., Harden, R., Howie, P., Ker, J. and Pippard, M. (2001). AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. *Medical Teacher*, 23(6), pp.535--551.

SECTION 2: CONCEPTS / TERMINOLOGY ‘HEADLINES’

This section includes expanded descriptions of key concepts relating to modern concepts of validity, standard setting, feedback and psychometrics.

Validity – Modern Concepts

Description: Validity addresses the question of whether a test measures what it is supposed to measure. While validity[1] has been considered as one of the characteristics of specific assessment instruments (van der Vleuten, 1996) together with reliability, educational impact, acceptability and cost, more modern concepts of validity are more detailed and encompassing. Hence in this updated guide, we adopt this concept and move from a notion of validity as it relates to the specific assessment instruments described. This overarching unitary view of validity was initially described and developed by Messick (1989, 1996) and further developed by Kane (2001, 2006) and Messick (2014). It is becoming increasingly accepted as a foundation stone for evaluating assessment tools or whole programmes of assessment (Schuwirth and van der Vleuten, 2012)

The fundamental concept of validity is whether the decisions made on the basis of particular tests can be reasonably defended. Therefore, there are certain criteria and evidence which need to be documented and presented to support the decisions which are made as a result of any test (examination). These criteria should be considered for every test which has summative impact on candidates' lives (i.e. high-stakes tests), such as progression from one year to another in undergraduate education, graduation, or certification for postgraduate degrees.

Considerations: The following criteria need to be considered as impacting on this overall unified concept of validity. Although listed separately, they are clearly linked and in some cases complementary and should be viewed as such.

1. Test content: Refers to the purpose of the test and how it is defined i.e. Does the test content appropriately reflect the learning objectives of the course/module? Blueprinting is key to this aspect of validity.

2. Response process: Refers to what kind of testing formats are being used i.e: chosen or constructed. Chosen is where the candidates choose from a list of answers offered within the test whereas constructed is where the candidates generate answers for themselves. There is more scope for error with constructed responses as these generally cannot be electronically scored and require examiners to read and mark. Certain criteria should therefore be met to minimize this error ie.:

- Clearly set out outlineanswers with scoring rubrics
- Blinded double marking
- A clear process for moderation where there is a difference in score between examiners

3. Internal structure: Refers to how a test is constructed and includes the following criteria:

- Number of items (in a written test) or numbers of stations (in a practical test)
- Format of the items
- Whether the format is appropriate for the domain of skill being tested (e.g. MCQs for knowledge tests, OSCEs for clinical skills)
- Is there sufficient sampling for the tests to be reliable?
- Scrutiny of the psychometric analyses of the test i.e. reliability coefficients
- Item analysis data (test score correlation, facility indices, etc)
- Are all parts of the test equally weighted?
- Is there compensation?
- What standard setting method is applied to determine the pass mark?

4. Relationship to other tests: How do the results of a test compare to the results of other tests taken by the same candidates?

5. Effects/outcomes: Consider the implications and consequences of decisions made on the basis of each test e.g.

- Effect on student learning
- Impact of failing - on students, on parents, on remediation and support staff
- Impact of passing - on students, on patients, animal welfare, client satisfaction, university reputation, regulatory body.

Key Points:
Modern concepts of validity are all-encompassing and do not consider validity as a property of an individual assessment instrument

- Evidence needs to be gathered against a range of criteria to ensure an overall programme of assessment is valid for the purposes intended

References and Further Reading

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), pp.319--342.

Kane, M. (2006). Validation. In: R. Brennan, ed., *Educational Measurement*, 1st ed. Westport, CT: Praeger, pp.7-64.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1996). Standards-based Score Interpretation: Establishing Valid Grounds for Valid Inferences. In: *Proceedings of the Joint Conference on Standard Setting for Large Scale Assessments, Sponsored by National Assessment Governing Board and the National Center for Education Statistics*. Washington, DC: Government Printing Office.

Messick, S. (2014). Validity. In: R. Linn, ed., *Educational Measurement*, 3rd ed. New York: Macmillan, pp.13-103.

Schuwirth, L. and van der Vleuten, C. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, 33(10), pp.783--797.

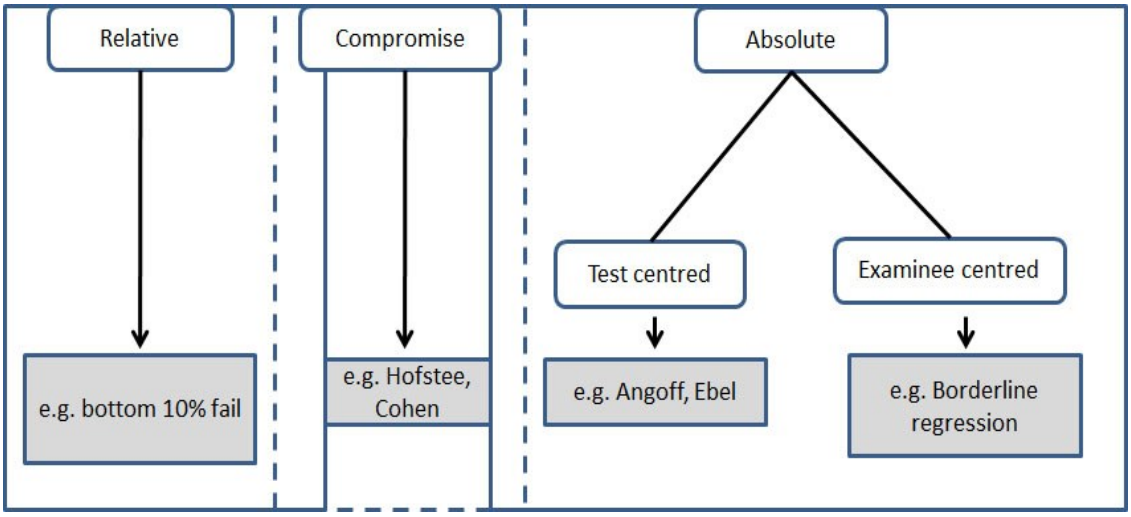
Schuwirth, L. and van der Vleuten, C. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, 46(1), pp.38--48.

Van der Vleuten, C. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*, 1(1), pp.41--67.

[1] For a more detailed breakdown of different types of validity which have been described, please see the Glossary

Standard Setting

Description: Standard setting is the process whereby decisions are made about boundaries or 'cut-points' between groups of students. Most commonly this decision focusses on those who pass and those who fail but the process can also be applied to other boundaries e.g. those who gain distinction or other form of credit and those who do not. Standards can be described as relative (norm referenced), absolute (criterion referenced) or compromise.



Relative Standards
The performance of candidates is reported relative to each other. Relative standards may be used for ranking of candidates e.g. for courses which may be competitive or in admissions.

Absolute Standards
A decision is made before the test is taken about the difficulty of the test and the requirements for passing. In theory, using absolute methods, all candidates could pass and all could fail. Such absolute standards are most appropriate for tests of competence when we want to be assured that candidates are 'safe' either to move to the next phase of the curriculum or out into practice. Absolute standards can be further considered as either 'test-centred' or 'examinee-centred'.

Test-Centred Absolute Methods
Two of the best known 'test-centred' methods for establishing an absolute standard on MCQ assessments are the Angoff and Ebel method. Both these methods rely on judges estimating the performance of a hypothetical group of 'borderline' candidates in the context of the assessment they are setting the standard for.

Examinee-Centred Absolute Methods
In these methods (which are common in OSCEs), the standard takes into account the performance of individual candidates based on overall criteria or overall test performance. A commonly used example is the borderline regression method where candidates are marked on a checklist and then given an additional overall global rating. The checklist score is then regressed on the global rating which then allows calculation of the checklist passing score.

Compromise Standards
In these methods, elements of both relative and absolute standards are incorporated. The best example of this is the Hofstee method where decisions are made in advance about the tolerance rates for failure and also the minimum and maximum acceptable cut-point for the given assessment.

Considerations: Although the rationale for absolute standard setting is clear, in practice several factors need to be considered:

1. Number of judges. Standard setting panels typically require 6-8 judges. Whilst this can be easy to achieve e.g. in a final exanimation of competence when staff may feel comfortable with the concept of a borderline student, at earlier stages in the curriculum or in small disciplines this may be more difficult to achieve. Where veterinary medicine has an additional complication factor is in the multispecies nature of the core curriculum.
2. It is well recognised that for judges, conceptualising the borderline student can be challenging
3. Where absolute methods are used, examination boards should ensure that the process produces a credible result and should have a well described pre-published strategy to deal with any anomalies

Key Points:

- Standards can be relative or absolute
- The method chosen should relate to the purpose of the assessment and should be defensible
- Standard setting is resource intensive and may, at least initially, be conceptually challenging
- Significant staff development needs to be implemented to ensure a robust standard setting process

References and Further Reading

Angoff, W. (1971). Scales, norms and equivalent scores. In: R. Thorndike, ed., *Educational Measurement*, 2nd ed. Washington, DC: American Council on Education, pp.508-600.

Bandaranayake, R. (2008). Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher*, 30(9-10), pp.836--845.

Boursicot, K., Roberts, T. and Pell, G. (2007). Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education*, 41(11), pp.1024--1031.

De Gruijter, D. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement*, 22(4), pp.263--269.

Downing, S., Tekian, A. and Yudkowsky, R. (2006). Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education. *Teaching and Learning in Medicine*, 18(1), pp.50--57.

Friedman Ben-David, M. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, 22(2), pp.120--130.

Kramer, A., Muijtjens, A., Jansen, K., Dusman, H., Tan, L. and van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education*, 37(2), pp.132--139.

Norcini, J. (2003). Setting standards on educational tests. *Medical Education*, 37(5), pp.464--469.

Reid, J. (1991). Training Judges to Generate Standard-Setting Data. *Educational Measure: Issues Practice*, 10(2), pp.11-14.

Feedback

Description

It is well recognised (for example in results of the UK National Students' Survey) that students in professional educational programmes often complain of a lack of feedback. On the other hand, teachers are frustrated that their efforts to provide feedback go unrecognised and do not seem to effect significant change. Whilst summative assessments such as examination give students an indication of their performance relative to their peers, there is an increasing drive for more effective formative feedback to become part of the culture. Traditionally feedback has been considered to be a teacher-led process whereby the student is given information about their performance. However modern approaches encourage a dialogue between trainer and trainee, with trainees taking a more active role in seeking and using feedback. Additionally, an effective feedback process should help develop the student's self-evaluation skills. Feedback in the preclinical environment might include, for example, discussion of a piece of coursework with a tutor, or peer feedback within a small group setting. Feedback in the clinical environment commonly takes the form of "in-the-moment" feedback during routine clinical work, regular progress discussions with a tutor, or written feedback at the end of a clinical placement. It is important that both positive and negative aspects of a students' performance are discussed in a timely, accurate, non-judgemental manner, using specific examples, and that the student is supported to engage with and act upon the feedback. Staff training in techniques for increasing the effectiveness of feedback, and student training in seeking and using feedback, can be invaluable in improving the "feedback culture" within the teaching environment.

Considerations / Practicalities

1. Feedback takes time. However, techniques such as the "One-minute teacher" (Neher and Stevens, 2003) are described for maximising teaching opportunities and feedback in clinics without disrupting a busy clinical workload.
2. Staff members need to prioritise feedback discussions with students. This requires a culture of feedback within the teaching environment, and training of staff to increase their confidence and skills in feedback dialogue. Feedback given in an inappropriate manner can be ineffective or indeed harmful.
3. Staff can find it hard to give critical feedback, particularly when it relates to issues of professionalism rather than knowledge or skills. Training, and an increased expectation of a feedback dialogue by students, may help to overcome this.
4. Students need to seek, recognise and act on feedback, which may require some proactive student training.

Key Points

- Feedback is essential for the effective development of professionals, and development of an effective "feedback culture" is paramount
- Written or verbal feedback should be timely, accurate, specific, objective, non-judgemental and balanced
- Every feedback interaction should generate a plan for the student's improvement

References and Further Reading

Boud, D. and Molloy E. (2013). Feedback in Higher and Professional Education: Understanding it and doing it well. Routledge, Oxon.

Neher, J. and Stevens, N. (2003). The one-minute preceptor: shaping the teaching conversation. *Family Medicine-Kansas City*, 35(6), pp.391--393.

Nicol, D. J. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education,* 31(2), pp.199-218.

Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education*. Assessment & Evaluation in Higher Education,* 35(5), pp.501-517.

Pendleton, D., Schofield, T., Tate, P., and Havelock, P. (2003) The consultation: An approach to learning and teaching: Oxford: Oxord University Press

Ramani, S. and Kracko, S. K. (2012) Twelve tips for giving feedback effectively in the clinical environment. *Medical Teacher,* 34, pp.787-791

Van de Ridder, J.M., Stokking, K.M., McGaghie, W.C. and ten Cate, O.T. (2008). What is feedback in clinical education? *Medical Education* 42, pp. 189-197

Psychometrics

Description: Psychometrics in this context refers to the application of statistical methods to assessment data to ensure that the assessment process is accurate i.e. reliable and valid. Validity is discussed above; here we will focus on reliability i.e. the reproducibility of the results. An assessment process cannot be valid if it is not reliable; however reliability does not guarantee validity (Hecker and Violato, 2009). Three main approaches to modelling responses to assessment have been developed (Schuwirth and van der Vleuten, 2011., McManus, 2010):

Classical test theory: This is the most widely used theory. It assumes that a candidate has a true ability (true score) but that the actual score is influenced by measurement errors. It is of most use in multiple choice tests when all students are given the same set of questions. Commonly analysed statistics include the P-value (the proportion of candidates answering the question correctly); the item-total correlation (the discriminatory power of the individual item), and Cronbach's α (the internal consistency of the test). It is generally considered that Cronbach's α should be >0.8 in a high stakes assessment, and that if it is >0.9 it is likely that there is some redundancy in the test (e.g. the test may contain more items than necessary for reliability or is repeatedly sampling the same knowledge base). For further discussion of the limitations of Cronbach's α see Schuwirth and van der Vleuten (2011) and Tavakol and Dennick (2011).

Generalisability theory: This is more useful when there is the potential for multiple sources of measurement error within an assessment e.g. clinical or OSCE style assessments where not all candidates are seen by all examiners, or may not all see the same patient. It can be used to identify variability due to different examiners (e.g. hawks and doves), and also allows the examining team to answer questions such as "How would the reliability be affected by having e.g. fewer stations or fewer examiners?"

Item-response theory: This requires large data sets and is best used for testing carried out at a large-scale level (e.g. national level testing). It calculates the difficulty of items as well as the discriminative value and the likelihood of chance-guessing; it estimates item difficulty and student ability independently of each other. It requires complex mathematical modelling and significant input from a psychometrician.

Considerations:

1. Psychometrics is a discipline in its own right; expert input from a psychometrician is extremely valuable.
2. When decisions are made around a particular score, e.g. a cut score for pass/fail decisions, it is valuable to calculate the standard error of measurement (SEM) in order to establish 95% confidence intervals around this cut score. (*Note that this then raises the issue that, for students whose scores fall within these confidence intervals, it is not possible to conclude (with a p≤0.05) whether or not these students have passed the test. Some assessment teams will then raise the pass mark to account for this uncertainty, in order to have confidence in the reliability of a passing score representing a true pass)*

Key Points:

- Psychometric principles are increasingly being adopted as a standard part of professional programme's assessment protocols to evaluate and continually refine/ improve assessments.
- Classical test theory is currently the most widely used and veterinary educators are increasingly developing expertise in its use.
- Familiarisation with the basics is straightforward, but users should also familiarise themselves with the limitations and underlying assumption in order to be confident in their interpretation of the results.

References and Further Reading

Hecker, K. and Violato, C. (2009). Validity, reliability, and defensibility of assessments in veterinary education. *Journal of Veterinary Medical Education*, 36(3), pp.271--275.

McManus, C. (2010). Focus on: The measurement of reliability. In: T. Swanick (Ed), ed., *Understanding Medical Education: Evidence, Theory and Practice*, 1st ed. Chichester: Wiley-Blackwell.

Schuwirth, L. and van der Vleuten, C. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, 33(10), pp.783--797.

Tavakol, M. and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, pp.53--55.

Glossary of Selected Terms

Blueprint: Indicates for an exam the content / areas covered, and relates to the learning objectives of the course. The relative amounts / approximate number of questions in each area can be defined.

Criterion referencing: Assessment is linked to achievement of outcomes regardless of the performance of other students i.e. measured against a defined criterion, absolute requirement or objective. Theoretically all students could pass or all could fail.

Cueing effects: In MCQs, and similar exam formats, examinees are able to eliminate wrong answers and recognise the correct answer, rather than needing to work out the answer. Questions should be designed to avoid cueing.

Formative Assessment: Sometimes referred to as 'assessment for learning' and provides information and feedback to the student on their performance. This allows the student to use the feedback to inform and guide future learning.

Global rating scales: These differ from checklists as the rater(s) assess each skill on a scale with categories that represent a range of performance e.g. from unsatisfactory to above expected performance levels. The forms usually include assessment of a range of skills such as: technical ability, consultation skills, knowledge, history taking, professionalism, team working and communication. They may include areas for the examiner to provide feedback comments. Global rating scales are used in a number of assessment methods e.g. OSCEs and mini-CEX.

Halo effects: Can be used to describe:
a) the effect whereby a judgement on one aspect is influenced by an overall impression of the person
b) the effect whereby a judgement is influenced by the performance of previous candidates in contrast to the current candidate i.e. after one or more consecutive poor candidates the subsequent candidate, even if average, would appear good and be 'over scored'.

Hawks and Doves: Two categories of examiners. Hawks tend to mark examinees 'down', while doves are lenient and award higher marks than the average across the board. When a hawk and a dove are together, the hawk tends to dominate.

Norm referencing: Refers to assessment which aims to discriminate between students by ranking them or 'grading on a curve'. The achievement of one student is relative to the rest of the students in that cohort.

Summative Assessment: Usually associated with a mark or grade and often occurs towards the end of a course. There is clearly overlap between these two categories as results and feedback from summative tests can also be used formatively.

Validity:

> **Face validity**: the assessment method, on first impression, appears to measure the intended competency
> **Content validity**: refers to the content of the assessment and how representative it is of the desired learning objectives. In practice, ensuring content validity typically involves the creation of a blueprint or spreadsheet to facilitate mapping of the assessment to the learning objectives.

Construct validity: refers to whether the scores on an assessment align with the trait the assessment is intended to measure
Criterion-related validity: refers to how well the assessment relates to some other criterion. This may be predictive (where the criterion of interest is future performance) or concurrent (where the criterion of interest is another criterion measured at the same time)
Consequential validity: refers to the impact the use of the assessment may have on student behaviour

REFERENCES

The following list combines all the references from the previous sections together with other reading material and references which the authors have found useful.

A North American Study of the Entry-Level Veterinary Practitioner. (2010). 1st ed. [ebook] Bismarck: National Board of Veterinary Medical Examiners. Available at: http://www.nbvme.org/image/cache/2010_NAVLE_job_analysis_report.pdf [Accessed 6 May. 2014].

Anderson, J. (2004). Multiple choice questions revisited. *Medical Teacher*, 26(2), pp.110--113.

Anderson, L. and Krathwohl, D. (2001). *A taxonomy for learning, teaching, and assessing*. 1st ed. New York: Longman.

Angoff, W. (1971). Scales, norms and equivalent scores. In: R. Thorndike, ed., *Educational Measurement*, 2nd ed. Washington, DC: American Council on Education, pp.508-600.

Anon, (2006). 1st ed. [ebook] Available at: http://www.veteducation.org/resources/view_workshop2006_SCT_examples.pdf [Accessed 6 May. 2014].

Assessment in undergraduate medical education. (2009). 1st ed. [ebook] General Medical Council. Available at: http://www.gmc-uk.org/Assessment_in_undergraduate_medical_education_0211.pdf_48902978.pdf [Accessed 7 May. 2014].

Bandaranayake, R. (2008). Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher*, 30(9-10), pp.836--845.

Beullens, J., Damme, B., Jaspaert, H. and Janssen, P. (2002). Are extended-matching multiple-choice items appropriate for a final test in medical education? *Medical Teacher*, 24(4), pp.390--395.

Bloom, B. (1984). Taxonomy of Educational Objectives. In: D. McKay, ed., *The Cognitive Domain*, 1st ed. New York: Company Inc.

Boud, D. and Molloy, E. (2013). *Feedback in higher and professional education*. 1st ed. London: Routledge.

Boursicot, K., Roberts, T. and Pell, G. (2007). Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education*, 41(11), pp.1024--1031.

Case, S. and Swanson, D. (2002). *Constructing Written Test Questions For the Basic and Clinical Sciences*. 3rd ed. [ebook] Philadelphia: National Board of Medical Examiners. Available at: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf [Accessed 6 May. 2014].

Challis, M. (1999). AMEE Medical Education Guide No. 11 (revised): Portfolio-based learning and assessment in medical education. *Medical Teacher*, 21(4), pp.370--386.

Charlin, B., Roy, L., Brailovsky, C., Goulet, F. and van der Vleuten, C. (2000). The Script Concordance test: a tool to assess the reflective clinician. *Teaching and Learning in Medicine*, 12(4), pp.189--195.

Crosby, J. (2003). Assessment of the Student Practitioner. In: J. Sweet, S. Huttly and I. Taylor, ed., *Effective Learning & Teaching in Medical, Dental & Veterinary Education*, 1st ed. London: Kogan Page, pp.71-89.

Cunnington, J., Hanna, E., Turnhbull, J., Kaigas, T. and Norman, G. (1997). Defensible assessment of the competency of the practicing physician. *Academic Medicine*, 72(1), pp.9--12.

Cunnington, J., Neville, A. and Norman, G. (1996). The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education*, 1(3), pp.227--233.

Davis, M., Harden, R., Howie, P., Ker, J. and Pippard, M. (2001). AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. *Medical Teacher*, 23(6), pp.535--551.

Davis, M. and Karunathilake, I. (2005). The place of the oral examination in today's assessment systems. *Medical Teacher*, 27(4), pp.294--297.

Davis, M. and Ponnamperuma, G. (2005). Portfolio assessment. *Journal of Veterinary Medical Education*, 32(3), pp.279-284.

Davis, M., Ponnamperuma, G., McAleer, S. and Dale, V. (2006). The objective structured clinical examination (OSCE) as a determinant of veterinary clinical skills. *Journal of Veterinary Medical Education*, 33(4), pp.578--587.

De Gruijter, D. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement*, 22(4), pp.263--269.

Dictionary.com, (2014). *Dictionary.com*. [online] Available at: http://dictionary.reference.com [Accessed 25 Jan. 2014].

Downing, S., Tekian, A. and Yudkowsky, R. (2006). Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education. *Teaching and Learning in Medicine*, 18(1), pp.50--57.

Evans, R., Elwyn, G. and Edwards, A. (2004). Review of instruments for peer assessment of physicians. *British Medical Journal*, 328(7450), p.1240.

Friedman Ben-David, M. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, 22(2), pp.120--130.

Gagnon, R., Charlin, B., Lambert, C., Carriere, B. and van der Vleuten, C. (2009). Script concordance testing: more cases or more questions? *Advances in Health Sciences Education*, 14(3), pp.367--375.

Guidance notes for CBD. (2010). 1st ed. [ebook] The Royal College of Surgeons of England. Available at: https://www.iscp.ac.uk/static/public/cbd_guidance.pdf [Accessed 6 May. 2014].

Harden, R., Stevenson, M., Downie, W. and Wilson, G. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1(5955), p.447.

Hecker, K. and Violato, C. (2009). Validity, reliability, and defensibility of assessments in veterinary education. *Journal of Veterinary Medical Education*, 36(3), pp.271--275.

Hecker, K., Read, E., Vallev, Krebs, G., Donszelmann, D., Muelling, C. and Freeman, S. (2010). Assessment of First-Year Veterinary Students' Clinical Skills Using Objective Structured Clinical Examinations. *Journal of Veterinary Medical Education*, 37(4), pp.395--402.

Hodges, B. and McIlroy, J. (2003). Analytic global OSCE ratings are sensitive to level of training. *Medical Education*, 37(11), pp.1012--1016.

Hodges, B. (2006). The objective structured clinical examination: Three decades of development. *Journal of Veterinary Medical Education*, 33(4), pp.571--577.

Hopkins, K. (1998). Test Validity. In: K. Hopkins, ed., *Educational and Psychological Measurement and Evaluation*, 8th ed. Needham Heights, MA: Allyn & Bacon.

Jennett, P. and Affleck, L. (1998). Chart audit and chart stimulated recall as methods of needs assessment in continuing professional health education. *Journal of Continuing Education in the Health Professions*, 18(3), pp.163--171.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), pp.319--342.

Kane, M. (2006). Validation. In: R. Brennan, ed., *Educational Measurement*, 1st ed. Westport, CT: Praeger, pp.7-64.

Kogan, J., Bellini, L. and Shea, J. (2002). Implementation of the Mini-CEX to Evaluate Medical Students' Clinical Skills. *Academic Medicine*, 77(11), pp.1156--1157.

Kramer, A., Muijtjens, A., Jansen, K., Dusman, H., Tan, L. and van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education*, 37(2), pp.132--139.

Ma, I., Zalunardo, N., Pachev, G., Beran, T., Brown, M., Hatala, R. and McLaughlin, K. (2012). Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. *Advances in Health Sciences Education*, 17(4), pp.457--470.

Magnier, K., Dale, V. and Pead, M. (2012). Workplace-Based Assessment Instruments in the Health Sciences. *Journal of Veterinary Medical Education*, 39(4), pp.389--395.

May, S. and Head, S. (2010). Assessment of technical skills: best practices. *Journal of Veterinary Medical Education*, 37(3), pp.258--265.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher*, 26(8), pp.709--712.

McLeod, R., Mires, G. and Ker, J. (2012). Direct observed procedural skills assessment in the undergraduate setting. *The Clinical Teacher*, 9(4), pp.228--232.

McManus, C. (2010). Focus on: The measurement of reliability. In: T. Swanick (Ed), ed., *Understanding Medical Education: Evidence, Theory and Practice*, 1st ed. Chichester: Wiley-Blackwell.

Messick, S. (1996). Standards-based Score Interpretation: Establishing Valid Grounds for Valid Inferences. In: *Proceedings of the Joint Conference on Standard Setting for Large Scale Assessments, Sponsored by National Assessment Governing Board and the national Center for Education Statistics*. Washington, DC: Government Printing Office.

Messick, S. (2014). Validity. In: R. Linn, ed., *Educational Measurement*, 3rd ed. New York: Macmillan, pp.13-103.

Meterissian, S., Zabolotny, B., Gagnon, R. and Charlin, B. (2007). Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *The American Journal of Surgery*, 193(2), pp.248--251.

Miller, G. (1990). The assessment of clinical skills/ competence/ performance. *Academic Medicine*, 65(9), pp.63--67.

Muzzin, L. and Hart, L. (1985). Oral Examinations. In: V. Neufeld and G. Norman, ed., *Assessing Clinical Competence*, 1st ed. New York: Springer Publishing Company, pp.71-93.

Neher, J. and Stevens, N. (2003). The one-minute preceptor: shaping the teaching conversation. *Family Medicine-Kansas City*, 35(6), pp.391--393.

Nicol, D. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), pp.199--218.

Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, 35(5), pp.501--517.

Norcini, J. and Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31.*Medical Teacher*, 29(9-10), pp.855--871.

Norcini, J. (2003). Setting standards on educational tests. *Medical Education*, 37(5), pp.464--469.

Norcini, J. (2005). The Mini Clinical Evaluation Exercise (mini-CEX). *The Clinical Teacher*, 2(1), pp.25--30.

Norman, G., van der Vleuten, C. and De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education*, 25(2), pp.119--126.

Pendleton, D., Schofield, T. and Tate, P. (2003). *The New Consultation: An approach to learning and teaching*. 1st ed. Oxford: Oxford University Press

Prescott-Clements, L., van der Vleuten, C., Schuwirth, L., Hurst, Y. and Rennie, J. (2008). Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). *Medical Education*, 42(5), pp.488--495.

Rademakers, J., ten Cate, T. and Bar, P. (2005). Progress testing with short answer questions. *Medical Teacher*, 27(7), pp.578--582.

Ramani, S. and Krackov, S. (2012). Twelve tips for giving feedback effectively in the clinical environment. *Medical Teacher*, 34(10), pp.787--791.

RCVS Awards, (2014). *OSCE stations*. [online] Available at: http://awardingbody.rcvs.org.uk/examinations/practical-examinations/osce-stations/ [Accessed 8 May. 2014].

Regehr, G., MacRae, H., Reznick, R. and Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, 73(9), pp.993--7.

Reid, J. (1991). Training Judges to Generate Standard-Setting Data. *Educational Measure: Issues Practice*, 10(2), pp.11-14.

Rhind, S., Baillie, S., Brown, F., Hammick, M. and Dozier, M. (2008). Assessing Competence in Veterinary Medical Education: Where's the Evidence? *Journal of Veterinary Medical Education*, 35(3), pp.407-411.

Rhind, S. (2006). Competence at graduation: implications for assessment. *Journal of Veterinary Medical Education*, 33(2), pp.172--175.

Schuwirth, L. and van der Vleuten, C. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, 38(9), pp.974--979.

Schuwirth, L. and van der Vleuten, C. (2010). How to design a useful test: the principles of assessment. *Understanding Medical Education: Evidence, Theory and Practice*, pp.241--254.

Schuwirth, L. and van der Vleuten, C. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, 33(10), pp.783--797.

Schuwirth, L. and van der Vleuten, C. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, 46(1), pp.38--48.

Schuwirth, L. (2003). ABC of learning and teaching in medicine: Written assessment. *British Medical Journal*, 326(7390), pp.643-645.

Schuwirth, L. (2007). The need for national licensing examinations. *Medical Education*, 41(11), pp.1022--1023.

Setna, Z., Jha, V., Boursicot, K. and Roberts, T. (2010). Evaluating the utility of workplace-based assessment tools for speciality training. *Best Practice & Research. Clinical Obstetrics & Gynaecology*, 24(6), pp.767--782.

Streiner, D. and Norman, G. (2003). Biases in responding. In: G. Norman and D. Streiner, ed., *Health Measurement Scales*, 3rd ed. Oxford: Oxford University Press.

Swing, s. and Bashook, P. (2000). *Toolbox of Assessment Methods*. 1st ed. [ebook] Accreditation Council for Graduate Medical Education and American Board of Medical Specialities. Available at: https://dconnect.acgme.org/Outcome/assess/Toolbox.pdf [Accessed 7 May. 2014].

Tavakol, M. and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, pp.53--55.

Tractenberg, R., Gushta, M., Mulroney, S. and Weissinger, P. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education*, 18(5), pp.945--961.

Turnbull, J. and Van Barneveld, C. (2002). Assessment of clinical performance: in-training evaluation. *Springer International Handbooks of Education*, 7, pp.793--810.

van Bruggen, L., Manrique-van Woudenbergh, M., Spierenburg, E. and Vos, J. (2012). Preferred question types for computer-based assessment of clinical reasoning: a literature study. *Perspectives on Medical Education*, 1(4), pp.162--171.

Van de Ridder, J., Stokking, K., McGaghie, W. and Ten Cate, O. (2008). What is feedback in clinical education? *Medical Education*, 42(2), pp.189--197.

Van der Vleuten, C. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*, 1(1), pp.41--67.

Van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *British Medical Journal*, 321(7270), p.1217.

Van der Vleuten, C., Norman, G. and De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education*, 25(2), pp.110--118.

van Mook, W., van Luijk, S., O'Sullivan, H., Wass, V., Schuwirth, L. and van der Vleuten, C. (2009). General considerations regarding assessment of professional behaviour. *European Journal of Internal Medicine*, 20(4), pp.90--95.

Wass, V., McGibbon, D. and Van der Vleuten, C. (2001). Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Medical Education*, 35(4), pp.326--330.

Wass, V., Wakeford, R., Neighbour, R. and van der Vleuten, C. (2003). Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. *Medical Education*, 37(2), pp.126--131.

Wilkinson, J., Crossley, J., Wragg, A., Mills, P., Cowan, G. and Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education*, 42(4), pp.364-373.

Wilkinson, T., Wade, W. and Knock, L. (2009). A blueprint to assess professionalism: results of a systematic review. *Academic Medicine*, 84(5), pp.551--558.

Wood, L., Hassell, A., Whitehouse, A., Bullock, A. and Wall, D. (2006). A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. *Medical Teacher*, 28(7), pp.185--191.

Wood, L., Wall, D., Bullock, A., Hassell, A., Whitehouse, A. and Campbell, I. (2006). 'Team observation': a six-year study 1 of the development and use of multi-source feedback (360-degree assessment) in obstetrics and gynaecology training in the UK. *Medical Teacher*, 28(7), pp.177--184.

Wragg, A., Wade, W., Fuller, G., Cowan, G. and Mills, P. (2003). Assessing the performance of specialist registrars. *Clinical Medicine*, 3(2), pp.131--134.

Acknowledgements

We are grateful to Professor Kathy Boursicot and Dr Claire Phillips for their helpful discussions and input. The support of the Higher Education Academy in printing this booklet is also gratefully acknowledged.